How can I group a data frame in dplyr by all but one column?

Authored by

stats writer

June 25, 2024

In dplyr, a popular data manipulation package in R, it is possible to group a data frame by all but one column using the "group_by" function. This allows for the creation of subgroups within the data frame based on the remaining columns, while keeping the specified column as a single group. This allows for efficient grouping and analysis of data sets with multiple variables, providing greater flexibility and insight into the data.

Group by All But One Column in dplyr

You can use the following basic syntax to group by all columns but one in a data frame using the package in R:

```
df %>%
group_by(across(c(-this_column)))
```

This particular example groups the data frame by all of the columns except the one called this_column.

Note that the negative sign (-) in the formula tells dplyr to exclude that particular column in the group_by() function.

The following example shows how to use this syntax in practice.

Example: Group by All But One Column in dplyr

Suppose we have the following data frame in R that

contains information about various basketball players:

```
#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
position=c('G', 'G', 'F', 'F', 'G', 'G', 'F', 'F'),
starter=c('Y', 'Y', 'Y', 'N', 'Y', 'N', 'N', 'N'),
points=c(99, 104, 119, 113))

#view data frame
df
```

```
team position starter points
1 A G Y 99
2 A G Y 104
3 A F Y 119
4 A F N 113
5 B G Y 99
6 B G N 104
7 B F N 119
8 B F N 113
```

Now suppose we would like to find the max value in the points column, grouped by every other column in the data frame.

We can use the following syntax to do so:

```
library(dplyr)#group by all columns except points column and find max points
df %>%
group_by(across(c(-points))) %>%
mutate(max_points = max(points))

# A tibble: 8 x 5
# Groups: team, position, starter
team position starter points max_points

1 A G Y 99 104
2 A G Y 104 104
3 A F Y 119 119
4 A F N 113 113
5 B G Y 99 99
6 B G N 104 104
7 B F N 119 119
8 B F N 113 119
```

From the output we can see:

The max points value for all players who had a team value of A, position value of G, and starter value of Y

was 104.The max points value for all players who had a team value of A, position value of F, and starter value of Y was 119.The max points value for all players who had a team value of A, position value of F, and starter value of N was 113.

And so on.

Note that we could also get the same result if we typed out every individual column name except points in the group_by() function:

```
library(dplyr)#group by all columns except points column and find max points
df %>%
group_by(across(c(team, position, starter))) %>%
mutate(max_points = max(points))

# A tibble: 8 x 5
# Groups: team, position, starter
team position starter points max_points

1 A G Y 99 104
2 A G Y 104 104
3 A F Y 119 119
```

4 A F N 113 113

5 B G Y 99 99

6 B G N 104 104

7 B F N 119 119

8 B F N 113 119

This matches the result from the previous example.

The following tutorials explain how to perform other common tasks using dplyr: